Journal of Biomedical Informatics **34,** 301–310 (2001)
doi:10.1006/jbin.2001.1029

Wendy W. Chapman,*,[1] Will Bridewell,† Paul Hanbury,* Gregory F. Cooper,*,† and Bruce G. Buchanan*,†

**Center for Biomedical Informatics and †Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania 15213*

Received May 29, 2001; published online May 9, 2002

Narrative reports in medical records contain a wealth of information that may augment structured data for managing patient information and predicting trends in diseases. Pertinent negatives are evident in text but are not usually indexed in structured databases. The objective of the study reported here was to test a simple algorithm for determining whether a finding or disease mentioned within narrative medical reports is present or absent. We developed a simple regular expression algorithm called NegEx that implements several phrases indicating negation, filters out sentences containing phrases that falsely appear to be negation phrases, and limits the scope of the negation phrases. We compared NegEx against a baseline algorithm that has a limited set of negation phrases and a simpler notion of scope. In a test of 1235 findings and diseases in 1000 sentences taken from discharge summaries indexed by physicians, NegEx had a specificity of 94.5% (versus 85.3% for the baseline), a positive predictive value of 84.5% (versus 68.4% for the baseline) while maintaining a reasonable sensitivity of 77.8% (versus 88.3% for the baseline). We conclude that with little implementation effort a simple regular expression algorithm for determining whether a finding or disease is absent can identify a large portion of the pertinent negatives from discharge summaries.

Key Words: text classification; pertinent negatives; negation; narrative medical reports; natural language processing; artificial intelligence.

[1]To whom correspondence and reprint requests should be addressed at Center for Biomedical Informatics, 8084 Forbes Tower, University of Pittsburgh, Pittsburgh, PA 15213. Fax: (412) 383-3072. E-mail: chapman@cbmi.upmc.edu.

1. INTRODUCTION

Much of the clinical information contained in patient medical records is in narrative form and therefore unavailable to automated systems that could improve patient care or further medical research. Clinical information described in narrative reports is also difficult for humans to access for clinical, teaching, or research purposes.

Researchers in information retrieval are creating effective methods for automatically indexing narrative clinical documents to facilitate searching on relevant terms [1–3]. Information retrieval techniques, however, do not generally discriminate between terms that are mentioned as being present and terms that are negated. In fact, most phrases indicating negation are stop words in information retrieval systems and are not even used for indexing. In clinical reports the presence of a term does not necessarily indicate the presence of the clinical condition represented by that term. In fact, many of the most frequently described findings and diseases in discharge summaries, radiology reports, history and physical exams, and other transcribed reports are denied in the patient [4]. Physicians often note that a particular disease can be ruled out or that a finding consistent with a suspected disease is absent. We use the term “pertinent negatives” to refer to findings and diseases explicitly or implicitly described as

1532-0464/01 $35.00

absent in a patient. Differentiating pertinent negatives from positive conditions in a clinical report is crucial to accurate indexing of the report.

Researchers in the medical language processing community have created methods for automatically extracting information contained in narrative reports for decision support [5], guideline implementation [6, 7], detection and management of epidemics [8], and identification of patients eligible for research studies [9]. Medical language processing (MLP) systems do determine whether the extracted information is negated, but MLP techniques for negating clinical findings in free text are often entwined in the MLP system and are not transferable.

In this paper we describe and test a computationally simple algorithm that could be implemented quickly and easily to determine whether an indexed term is negated.

1.1. Negation in Natural Language

Whereas negation in predicate logic is well defined and syntactically simple, negation in natural language is complex and has been philosophized about for hundreds of years. Aristotle's theory of negation has its roots within his system of oppositions between pairs of terms. He described four species of opposition [10] including correlation (e.g., double vs half), contrariety (e.g., good vs bad), privation (e.g., blind vs sighted), and contradiction (e.g., He sits vs He does not sit). Pertinent negatives in clinical reports belong to the category of contradiction in which a proposition (e.g., experiencing nausea) is denied (i.e., not experiencing nausea). Identifying pertinent negatives, then, involves identifying a proposition ascribing a clinical condition to a person and determining whether the proposition is denied or negated in the text.

1.2. Previous Work on Negation

Whereas much has been published on negation in natural language [10, 14, 15], very little has been published on computational negation methods. McQuire and Eastman describe a method for disambiguating natural language queries to an information retrieval system. Their system detects ambiguous queries involving the negation phrase "not" and asks the user for clarification [16]. MLP systems described in the literature [17] perform syntactic and semantic processing to extract features from text, and features are augmented with information about uncertainty and negation. Phrases indicating negation are sometimes published in literature describing these systems (e.g. [18]), but the algorithms used to determine the scope of the negation phrases are entwined in the MLP system.

The most extensive study on negation was recently published by Mutalik *et al.* [19]. They use a lexical scanner with regular expressions and a parser that uses a restricted context-free grammar to identify pertinent negatives in discharge summaries and surgical notes. Like the algorithm described in this paper, their system first identifies propositions or concepts and then determines whether the concepts are negated. Their system performed with a sensitivity of 95.7% and a specificity of 91.8% and is fine tuned with rules that apply to particular negation phrases and syntactic structures. Mutalik's algorithm is quite complex and requires other utilities such as lex and Yacc. The tools required by Mutalik's algorithm are easily attainable; however, implementing their system in a preexisting indexing tool would be less straightforward than the regular expression algorithm we describe here.

1.3. Identifying Pertinent Negatives from Narrative Clinical Reports

The Unified Medical Language System (UMLS) provides a helpful resource for identifying propositions or concepts useful for medical indexing [11]. MEDLINE indexing uses sophisticated syntactic and semantic processing techniques, but does not incorporate explicit distinctions between positive and negative terms [20]. Various methods exist for indexing documents with UMLS phrases (see [12] for a good overview). Once a UMLS or other relevant concept has been marked in a clinical report, a separate negation algorithm could determine whether the concept is negated.

Accurate identification of pertinent negatives is not a simple problem since natural language is largely unstructured and allows great freedom of description. Hundreds of different phrases can be used to indicate denial of a finding or disease. Even if all possible phrases indicating negation could be identified, an algorithm must determine which propositions fall within the scope of the negation phrase. Consider the following sentence:

> "The chest X-ray showed no infiltrates and EKG revealed sinus tachycardia."

The negation phrase "no" applies to infiltrates but not to sinus tachycardia. Accurate analysis of scope may involve lexical, syntactic, or even semantic analyses.

In spite of the complexity of negation, we believe that in medical text negating clinical concepts is more restricted and therefore may not require full natural language understanding. We rely on the fact that medical narrative is a

sublanguage limited in its purpose. Quantitative studies have shown that medical documents are lexically less ambiguous than unrestricted documents [13]. We conjecture that pertinent negatives being described in the narrative reports are limited to a handful of semantic types, including findings, diseases, tests, drugs, etc., which are most often noun phrases rather than verbs, clauses, or sentences. Moreover, a few phrases constitute the majority of pertinent negations in different types of medical narratives [4]. Therefore, we believe that a simple negation algorithm like the one we describe below can accurately identify a large portion of pertinent negatives in medical narratives without utilizing sophisticated linguistic methodologies.

The algorithm we describe below is a simple algorithm that can be implemented quickly and easily into any medical concept indexing or feature extraction system.

2. METHODS

2.1. Algorithm

We designed an algorithm called NegEx to determine whether findings and diseases indexed from discharge summaries were negated by the dictating physician. To test our hypothesis that a relatively simple algorithm could produce reasonably accurate results, we compared NegEx against a baseline algorithm.

The input to NegEx is a sentence with indexed findings and diseases. The output is whether an indexed phrase is negated in the sentence. For this study, we automatically preprocessed the sentences, indexed the relevant phrases with UMLS terms, applied the negation algorithms, and compared the negations made by the algorithms against negations made by physicians. For NegEx to work, the indexing algorithm must first identify a UMLS term; when the indexing algorithm does not identify a UMLS term, a pertinent negative cannot be found.

We preprocessed the reports so that exactly one sentence appeared per line. Processing the reports by individual sentences means that information across sentences is not used in determining whether a clinical condition is affirmed or denied. Next, we removed all punctuation; information about syntactic structure, such as comma-delimited lists, is not used by NegEx. We did not remove stop words because some commonly used stop words (e.g., "of") are important parts of the expressions we look for. Finally, we indexed findings and diseases within the sentence by replacing phrases in the text with unique string identifiers from the UMLS.

Because the scope of the UMLS [11, 21] is purposefully broad, we chose an abridged set of phrases with a focus on diseases and findings. In particular, our focus was limited to the intersection of phrases in the International Statistical Classification of Disease and Related Health Problems, 10th revision (ICD10) [22] and phrases with the UMLS semantic type of "Finding," "Disease or Syndrome," or "Mental or Behavioral Dysfunction." Using a simple, string-matching program, each UMLS phrase in every sentence was automatically marked by replacing the phrase with its corresponding UMLS string ID. For example, the sentence

"The patient denied experiencing chest pain on exertion"

was rewritten as

"The patient denied experiencing ⟨S1459038⟩ on exertion."

Our string matching algorithm matched the longest possible string among eligible matches in the UMLS (i.e., "nonspecific viral rash" instead of "rash").

Both NegEx and a baseline algorithm were applied to the preprocessed sentences. The baseline algorithm is the algorithm previously in use by a system called the IPS system [9, 23] that was created at the University of Pittsburgh to help researchers identify relevant subsets of patient reports for research studies (we have since replaced the baseline algorithm with NegEx). The baseline algorithm searches for six phrases (see Appendix) that might indicate a negation and negates all UMLS terms following the negation phrase until the end of the sentence. The baseline algorithm is simplistic but had been used by the IPS system for 2 years. Informally tested on approximately 1000 sentences, the baseline algorithm was being used in a system where a human user reviews results and can filter out mistakes made by the algorithm.

NegEx expands on the baseline algorithm with additional negation phrases and a richer regular expression syntax yet retains much of the simplicity of the baseline method. Through a combination of manual scanning and semiautomated learning we identified 35 negation phrases (see Appendix) that could be divided into two groups. The first group (I), which we call "pseudo-negation" phrases, consists of phrases that appear to indicate negation but instead identify double negatives ("not ruled out"), modified meanings ("gram-negative"), and ambiguous phrasing ("unremarkable"). The second group (II) consists of phrases we believed are used to deny findings and diseases when used in one of

two regular expressions. In the first regular expression (II-A) the negation phrase precedes the UMLS term:

⟨negation phrase⟩ * ⟨UMLS term⟩.

In the second (II-B) the negation phrase follows the UMLS term:

⟨UMLS term⟩ * ⟨negation phrase⟩.

In both II-A and II-B the asterisk indicates that up to five tokens (i.e., words or UMLS terms) may fall between the negation phrase and the UMLS term.

The regular expressions were matched to the longest possible subset of the sentence. For example, the sentence "extremities showed *no* **cyanosis**, clubbing, or **edema**" (UMLS terms in bold, negation phrase in italics) matches the regular expression ⟨no⟩ * ⟨UMLS term⟩ twice with both "cyanosis" and "edema" being labeled as negated.

2.2. Training and Test Sets

The data used in this study were sentences from 2060 randomly selected, deidentified discharge summaries dictated between January 1, 1993 and December 31, 1995 at two medical ICU's at the University of Pittsburgh Medical Center. An arbitrarily selected subset of approximately 1500 reports from this dataset was extracted and used as a training set. We manually read the reports in the training set to determine what negation phrases and regular expressions would accurately identify pertinent negatives.

A separate test set was selected from the remaining 560 reports. To test the accuracy of our algorithm, we initially extracted all sentences from the set that contained UMLS terms of interest, namely UMLS terms also contained in ICD10 with one of the three previously mentioned semantic types. We then used a string-matching program to divide the sentences into two groups. The first group contained 500 sentences in which at least one of the negation phrases used by NegEx (lists I, II-A, and II-B in the Appendix) occurred. The main purpose of group 1 was to test the precision of NegEx's regular expressions. We also included in group 1 sentences containing phrases that we did not label as negation phrases but that we suspected might sometimes be used to signal a pertinent negative ("minimal sign of," "nonfocal," "nonspecific," and "unremarkable," "failed," "negative," "never," "nor," and "unable"). Approximately 15% of the reports from which we selected test sentences contained one of the phrases described in this paragraph.

The second group contained 500 sentences in which none of NegEx's negation phrases occurred. Group 2 was designed to determine the completeness of NegEx's set of negation phrases by capturing sentences with negation phrases we had not included. The entire test set contained a total of 1000 sentences.

2.3. Gold Standard

Three physicians judged the sentences in the test set to establish a "gold standard" against which the computerized algorithms could be compared. The physicians read overlapping subsets of the 1000 sentences and marked all UMLS terms in the sentences as (a) present—the term was described by the dictating physician as being present at the current visit, (b) absent—the term was explicitly described as being absent at the current visit, or (c) ambiguous—whether the term was present or absent was not clear from the sentence.

Each physician judged 400 of the 1000 sentences. To determine interrater reliability, 200 sentences were rated by pairs of physicians. Raters 1 and 2 judged 100 overlapping sentences, and raters 2 and 3 judged a separate set of 100 overlapping sentences. If two physicians judged the same sentence and did not agree on whether a UMLS term was present or absent, the term was marked as ambiguous.

2.4. Evaluation Techniques and Measures

Once human raters judged all terms, both the baseline algorithm and NegEx were applied to the test set. Whereas the raters judged each occurrence of a term in a sentence, NegEx treated multiple occurrences of a term in a sentence as a single occurrence. That is, if a term was used twice in one sentence, NegEx considered both occurrences of the term negated if at least one occurrence was negated in the sentence. For example, consider the sentence "the patient was placed under **neutropenic** precautions, and two days later, the patient was *no* longer **neutropenic**." A human rater might mark the first occurrence of neutropenic as positive and the second as negative. NegEx, however, would just report that neutropenic was negated in the sentence. To compensate for the different negating schemes, the raters' labels were manually examined. If a UMLS phrase was rated positive, negative, or ambiguous in all occurrences in a sentence, then no changes were made. If there was a disagreement (e.g., the first occurrence was rated positive and the second occurrence was rated negative), then both occurrences of the term were considered ambiguous in that sentence.

For both the baseline algorithm and NegEx sensitivity, specificity, positive predictive value (PPV), and negative

predictive value (NPV) were calculated for every UMLS term in the test set. True positive, true negative, false positive, and false negative counts were calculated as follows: True positive—NegEx and the rater negated the term; True negative—NegEx and the rater did not negate the term; False positive—NegEx negated the term but the rater did not negate the term; False negative—NegEx did not negate the term but the rater negated the term.

In addition, we analyzed the performance of the individual negation phrases and examined the accuracy of negating different UMLS phrases. Finally, we analyzed NegEx's errors to determine why the algorithm produced false positives and false negatives.

3. RESULTS

The 1000 test sentences contained 1235 occurrences of UMLS terms, 245 of which were unique strings. We examined interrater reliability of the physicians' judgments of the UMLS terms to evaluate the assumption that a reliable gold standard could be established from the judgment of a single physician. Of the 200 sentences that were rated by more than one physician, only a single term was rated differently. For the gold standard, that term was listed as ambiguous.

Table 1 describes the gold standard ratings produced by the physicians. The columns show the gold standard distribution of findings and diseases judged to be present, absent, and ambiguous in the two groups of test sentences and overall. The first column represents sentences in Group 1 (i.e., containing NegEx's negation phrases (lists II-A and II-B in the Appendix)), and the second column represents sentences in Group 2 (i.e., not containing NegEx's negation phrases). The last column contains the totals over all 1000 sentences.

TABLE 1
Number of Findings and Diseases in the Test Set That Were Given Ratings of Present, Absent, and Ambiguous by Physician Raters

Gold standard rating	Group 1 sentences (i.e., containing NegEx negation phrases) (n = 500)	Group 2 sentences (i.e., not containing NegEx negation phrases) (n = 500)	All sentences (n = 1000)
Present	186	542	728
Absent	324	19	343
Ambiguous	94	70	164
Total	604	631	1235

TABLE 2
Performance of the Baseline Algorithm

	Group 1 sentences (i.e., containing NegEx negation phrases) (n = 500) (%)	Group 2 sentences (i.e., not containing NegEx negation phrases) (n = 500) (%)	All sentences (n = 1000) (%)
Sensitivity	88.27	0.00	88.27
Specificity	52.69	100.00	85.27
PPV	68.42	—	68.42
NPV	79.46	96.99	93.01

Note. Sensitivity = Number of terms NegEx correctly negated/number of terms negated by rater; Specificity = Number of terms NegEx correctly did not negate/number of terms not negated by rater; PPV = Number of terms NegEx correctly negated/number of terms NegEx negated; NPV = Number of terms NegEx correctly did not negate/number of terms NegEx did not negate. Sensitivity is 0% in group 2 because group 2 did not contain sentences with negation phrases used by NegEx.

Tables 2 and 3 present performance statistics of both the baseline algorithm and NegEx, respectively. Terms that were listed by the gold standard as "explicitly absent" at the time of the patient's current visit were counted as the major category of interest, contrasted with the complement set that included both "present" and "ambiguous" (physician raters had to select from these three categories only and use their best judgment in classifying UMLS terms that did not fit cleanly into these categories, such as findings that were absent in the patient's past history). Since our focus was on NegEx's ability to identify pertinent negatives at the time of the current visit, we combined gold standard ratings other than *absent* (i.e., *present* and *ambiguous*) into a category of *not absent*. As can be seen from the tables, both the baseline

TABLE 3
Performance of NegEx

	Group 1 sentences (i.e., containing NegEx negation phrases) (n = 500) (%)	Group 2 sentences (i.e., not containing NegEx negation phrases) (n = 500) (%)	All sentences (n = 1000) (%)
Sensitivity	82.41	0.00	77.84
Specificity	82.50	100.00	94.51
PPV	84.49	—	84.49
NPV	80.21	96.99	91.73

algorithm and NegEx performed well. The baseline algorithm was more sensitive, whereas NegEx was more specific and precise.

Table 4 shows the 15 negation phrases actually identified by NegEx in the test set, along with their respective PPVs. Ten negation phrases shown in the Appendix did not occur in the test set at all. Of the 15 negation phrases found in the test set, three ("no," "without," and "no evidence of") accounted for 82% of the correctly identified pertinent negatives. Three of the 15 phrases had very poor PPV: "versus" (0%), "not" (58%), and "doubt" (50%) and are therefore candidates for further examination.

We evaluated NegEx's accuracy on the 28 UMLS terms that occurred in the test set at least 10 times, as shown in Table 5. Three terms (bleeding, seizure, and hepatitis) produced several false negatives (i.e., were not identified as pertinent negatives by NegEx but were by the gold standard raters). Most false negatives regarding bleeding and seizure were due to including "no further," "without further," and "without any further" in the list of pseudo-negation phrases (list I in the Appendix) rather than in the list of actual negation phrases (list II-A in the Appendix). The term hepatitis produced 7 false negatives that are interesting to consider. All 7 false negatives resembled the following sentences:

"The patient was Hepatitis A negative" or
"She received a Hepatitis vaccine."

In both sentences hepatitis is part of a string describing a test or vaccine. In both cases, hepatitis is absent in the patient. However, in the first sentence absence of the diseases is implied by negative results of the test. In the second sentence absence of the disease is implied by the presence of the vaccine.

Three terms produced several false positives (edema, infection, and pneumonia). There was no pattern in the causes of false positives for edema. Most of the false positives for infection and pneumonia were due to the negation phrase being used. Pneumonia's false positives resulted from the phrase "versus," and infection's from phrases of the form "no/not the source of infection."

TABLE 4
Positive Predictive Value (PPV) of NegEx Negation Phrases Found in the Test Set

Negation phrase	PPV (%)	Number of times identified by NegEx in test set
no signs of	100.00	2
ruled out	100.00	3
unlikely	100.00	2
absence of	100.00	1
not demonstrated	100.00	1
denies	100.00	7
no sign of	100.00	2
no evidence of	94.59	37
no	92.36	157
denied	90.00	10
without	88.10	42
negative for	80.00	10
not	58.33	24
doubt	50.00	2
versus	0.00	16

TABLE 5
Mistakes Made When Negating Frequently Occurring (i.e., at Least 10 Occurrences) UMLS Terms

UMLS Term	Number of occurrences	Number of false positives	Number of false negatives
bleeding	70	5	**12**
seizure	58	2	**8**
hepatitis	18	2	**7**
infection	49	5	3
pneumonia	67	5	2
edema	65	5	1
pulmonary edema	18	1	1
atelectasis	17	1	0
atrial fibrillation	17	1	0
cva	11	1	0
pneumothorax	26	1	0
thrombosis	16	0	2
stroke	10	0	1
ascites	12	0	1
hematuria	15	0	1
acidosis	15	0	1
congestive heart failure	23	0	1
weakness	26	0	1
hypertension	45	0	1
cyanosis	27	0	0
hemorrhage	10	0	0
hematemesis	11	0	0
respiratory distress	13	0	0
alcohol abuse	17	0	0
diabetes mellitus	21	0	0
esophagitis	11	0	0
sepsis	10	0	0
depression	10	0	0

Note. Terms in bold produced several false negatives, terms in italics produced several false positives.

4. DISCUSSION

Our results indicate that a simple regular expression algorithm using a small number of negation phrases performs

well at identifying pertinent negatives, in particular whether UMLS terms for findings and diseases indexed from discharge summaries are negated. When comparing the results of NegEx to those of the baseline, there is an obvious improvement in specificity and positive predictive value while maintaining reasonable sensitivity and negative predictive value. The main reason for improvement in positive predictive value is the limitation of a negation phrase's scope. The baseline algorithm negates anything following the negation phrase (up to the end of the sentence), whereas NegEx limited the number of words between the negation phrase and the UMLS term, enabling more accurate identification of the negation phrase's target. In addition, NegEx allows for more precision about the relative placement of the UMLS term to the negation phrase.

NegEx performs slightly worse than the baseline in sensitivity and negative predictive value. One reason for lower sensitivity is that limiting the number of words allowed between a negation phrase and the UMLS term fails to identify negated UMLS terms that are described in long lists. It is common for a physician to negate several findings or diseases in a comma-separated list. Some lists extend beyond the regular expression's word limit of five, so some of the terms in the list are not negated when they should be. A possible modification to NegEx would involve automatically identifying lists of UMLS terms and dynamically expanding the scope of the negation phrase to include the whole list. Additionally, identifying and incorporating more negation phrases could also improve sensitivity.

Table 4 shows that three negation phrases demonstrated poor sensitivity. Two phrases included as negation phrases are not actually used to negate clinical conditions but to indicate uncertainty about those conditions. "Versus" displayed 0% sensitivity over 16 terms. The sentence fragment ". . . pneumonia versus bronchitis for her cough" indicates uncertainty about whether the patient's cough is attributed to pneumonia or bronchitis. This interpretation was not clear to us when we selected the negation phrases from the training set, and as a result NegEx's positive predictive value and specificity were lowered. Had "versus" not been considered a negation phrase, NegEx would have displayed 96% specificity and 89% positive predictive value. "Doubt" displayed 50% sensitivity, but because "doubt" only triggered the negation algorithm two times in the test set, we are unable to draw meaningful conclusions about its performance. It is clear from both phrases, though, that a limitation of NegEx is only assigning a binary value of *absent*/*not absent* to the indexed terms. A more complete negation processor would also model uncertainty of indexed findings and diseases.

The third poorly performing phrase was "not" which received only 58% sensitivity over 24 occurrences. This finding is consistent with that of Mutalik *et al.* [19] and our previous report [4]. Determining the scope of "not" is complex. For example, the sentence

"This is not an infection"

indicates the clinical finding "infection" is absent. The "not" in the sentence

"This is not the source of the infection"

negates the term "source" but not the clinical finding "infection." World knowledge a native English speaker uses to determine the scope of "not" in the above examples would be difficult to represent in a simple algorithm. Consider also the following example:

"We did not treat the infection" and

"We did not detect an infection."

Although the sentences have similar syntactic structures, the finding "infection" is present in the patient in the former sentence and absent in the patient in the latter sentence. A more accurate negation algorithm would require further research into resolving the scope of the negation phrase "not."

A few pseudo-negation phrases actually acted as true negation phrases and should be moved from list I in the Appendix to list II-A. Specifically, "no further," "without further," and "without any further" were usually judged by the gold standard raters to signify pertinent negatives.

Other mistakes made by NegEx included missed negations (false negatives) due to passive syntactic structures indicating negation (e.g., "**nephrotic syndrome** was ruled out"), different negation phrases that were not included in our list (e.g., "a chest X-ray at this time was unremarkable for **pneumonia**"), and extensive modifiers between the negation phrase and the UMLS term (e.g., "no signs or symptoms of reoccurrence of his GI **bleeding**"). Falsely negated terms (false positives) were often caused by failure to decrease the scope of the negation phrase (e.g., "no **cyanosis** and positive **edema**") and by failure to distinguish current visits from the patient's past history (e.g., "no history of previous **cva**").

4.1. Limitations

One limitation in our methodology is using string matching to identify relevant UMLS phrases. Various methods

exist for indexing documents with UMLS phrases [12]. Using simple string matching decreases our ability to identify relevant UMLS phrases that might be matched with more sophisticated methods but reduces the noise created by false matches.

A major limitation in our study design was allowing the human raters to judge a term to be ambiguous. Because NegEx does not label a term as ambiguous, designing a method to compare NegEx's answers with the gold standard raters' answers in a fair way was difficult. If we ignored the terms judged ambiguous by the gold standard, we would have biased the results in our favor by getting rid of the difficult sentences. We chose to combine ambiguous judgments with positive judgments into a *not absent* complement set. Combining the two ratings caused the specificity to be higher than it might have been but gave lower values for sensitivity and PPV. Therefore, the sensitivity and PPV listed in this paper are lower than they might have been if we had not allowed an ambiguous rating.

A problem with NegEx's algorithm as described in this paper is that multiple occurrences of a UMLS term in a sentence are considered as one occurrence. Medical reports describe temporal relations among clinical concepts. Therefore, treating every occurrence of a concept individually will allow NegEx a better chance at accurately identifying pertinent negatives. For example, in this study the sentence "the patient was placed under **neutropenic** precautions, and two days later, the patient was *no* longer **neutropenic**" contributed one false positive by labeling neutropenic as negative in this sentence. The temporal relation between the two occurrences of neutropenic make it clear that the two occurrences should be labeled individually, in which case NegEx would label the first occurrence positive and the second negative.

A fundamental assumption of our approach is that a sentence-level analysis is sufficient for identifying pertinent negatives. For the most part we believe this simplifying assumption is reasonable. NegEx will probably miss some pertinent negatives because the UMLS term is referred to in another sentence by a pronoun such as "it" or a generic description of the term such as "the finding." We would capture more pertinent negatives by incorporating coreference resolution, but the amount of sophisticated processing needed would be substantial.

4.2. Future Work

Negation phrases appear to comply qualitatively with Zipf's law regarding the frequency distribution of words in human languages. From Zipf's law it follows that there are a few very common words, a middling number of medium-frequency words, and many low-frequency words [24]. Our results show that there are a few very common negation phrases ("no," "without, and "no evidence of"), more medium-frequency negation phrases, and a potentially huge number of low-frequency phrases. Therefore, including a few very common negation phrases can capture a large portion of the pertinent negatives. However, to increase the number of pertinent negatives identified by NegEx, we will continue adding and testing new negation phrases.

Along with testing new negation phrases, we plan a series of other experiments to expand NegEx's capabilities. First, we will test NegEx on an expanded set of UMLS terms. Although we limited our terms to those included in ICD10, the only terms NegEx had difficulty negating were diseases that are substrings of test or vaccine names (i.e., hepatitis). Therefore, we believe NegEx will generalize over the remaining terms in the UMLS semantic types examined in this study as well as over terms contained in other similar UMLS semantic types, i.e., semantic types that effectively represent findings and diseases (e.g., virus, symptom). Second, we will test NegEx on different report types. Because discharge summaries contain information from history and physical exams, lab reports, radiology exams, and other types of patient reports, we believe that discharge summaries are representative of other report types at describing findings and diseases and therefore the phrases listed in the Appendix will require little change. Third, to make NegEx more useful to general indexing and retrieval systems, we will also test its performance on different semantic types described in clinical documents, such as lab tests and medications. We believe that the regular expressions will likely remain the same but that different negation phrases will need to be added for different semantic types, but we must first test this hypothesis.

We also plan to expand NegEx's classification categories from *absent* versus *not absent at the current visit* to a more realistic representation of the temporal information described in reports. Medical reports describe findings and diseases in the past history of the patient, describe current findings with uncertainty, and prescribe plans based on findings and diseases that might occur in the future. To better characterize pertinent negatives NegEx needs to discriminate among past, present, and future concepts and to deal with uncertainty in the language describing the clinical concepts.

5. CONCLUSION

Our results indicate that a simple regular expression algorithm can accurately detect a large portion of pertinent negatives in discharge summaries. Identifying a set of pseudo-negation phrases, a set of negation phrases, and two simple regular expressions is all that is needed to identify most pertinent negatives in narrative medical records. The set of negation phrases and regular expressions presented here can be easily modified based on the results of this study. The phrases are not yet complete but are complete enough to achieve high levels of sensitivity and specificity while being easily implementable.

Note from the authors: A current and updated list of regular expressions and negation phrases used by NegEx can be found at http://omega.cbmi.upmc.edu/~chapman/NegEx.html.

APPENDIX

Negation phrases used in baseline algorithm

no	denies
not	without
*n't	ruled out
denied	

Negation phrases used in NegEx

I. Pseudo-negation phrases (false triggers, ambiguous negations, or double negatives)

- gram negative
- no further
- not able to be
- not certain if
- not certain whether
- not necessarily
- not rule out
- without any further
- without difficulty
- without further

II-A. Phrases used in regular expressions of the form ⟨phrase⟩ * ⟨UMLS term⟩ to indicate pertinent negatives, where the asterisk indicates 0–5 intervening words:

absence of	doubt
declined	negative for
denied	no
denies	no cause of
denying	no complaints of
did not exhibit	no evidence of
no sign of	versus
no signs of	without
not	without indication of
not demonstrate	without sign of
patient was not	ruled out
rules out	

II-B. Phrases used in regular expressions of the form ⟨UMLS term⟩ * ⟨phrase⟩ to indicate pertinent negatives, where the asterisk indicates 0–5 intervening words:

- declined
- unlikely

ACKNOWLEDGMENTS

The baseline algorithm was implemented by Paul Hanbury. NegEx was designed and implemented by Will Bridewell who also ran it to generate the results presented here. The lists of negation phrases were the result of group discussion. We thank Mehmet Kayaalp, M.D., and Lou Penrod, M.D., who judged the discharge summaries. (The third physician–reader was one of the coauthors, G.F.C.) This project was supported in part by National Library of Medicine Grants LM06625, LM07059, and LM006759.

REFERENCES

1. Hersh WR. Information retrieval: a health care perspective. New York: Springer-Verlag, 1996.
2. Aronson A, Rindflesch T, Browne A. Exploiting a large thesaurus for information retrieval. Proc RIAO '94 Conf 1994: 197–216.
3. Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. Proc AMIA Symp 1994; 240–4.
4. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp 2001; 105–9.
5. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000; 7:593–604.
6. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a

severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp 1999; 256–60.

7. Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp 2000; 235–9.
8. Hripcsak G, Knirsch CA, Jain NL, Stazesky RC, Pablos-mendez A, Fulmer T. A health information network for managing innercity tuberculosis: bridging clinical care, public health, and home care. Comput Biomed Res 1999; 32:67–76.
9. Aronis JM, Cooper GF, Kayaalp M, Buchanan BG. Identifying patient subgroups with simple Bayes'. Proc AMIA Symp 1999; 658–62.
10. Horn LR. A natural history of negation. Chicago, IL: Univ. of Chicago Press, 1989.
11. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993; 32:281–91.
12. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. J Am Med Inform Assoc 2001; 8:80–91.
13. Ruch P, Baud R, Geissbuhler A, Rassinoux AM. Comparing general and medical texts for information retrieval based on natural language processing: an inquiry into lexical disambiguation. Medinfo 2001; 10:261–5.
14. Horn L, Kato Y. Negation and polarity: syntactic and semantic perspectives. New York: Oxford Univ. Press, 1999.
15. Tottie G. Negation in English speech and writing: a study in variation. San Diego: Academic Press, 1991.
16. McQuire AR, Eastman CM. The ambiguity of negation in natural language queries to information retrieval systems. J Am Soc Inf Sci 1998; 49:686–92.
17. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999; 74:890–5.
18. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994; 1:161–74.
19. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001; 8:598–609.
20. Personal correspondence between B.G.B. and Alexa McCray.
21. UMLS knowledge sources—documentation. National Library of Medicine, 1997.
22. International statistical classification of disease and related health problems, 10th revision. Geneva: World Health Organization, 1998.
23. Cooper GF, Buchanan BG, Kayaalp M, Saul M, Vries JK. Using computer modeling to help identify patient subgroups in clinical data repositories. Proc AMIA Symp 1998; 180–4.
24. Manning CD, Schutze H. Foundations of statistical natural language processing. Cambridge, MA: MIT Press, 1999.